Constraint-based modeling and machine learning applications for analysis and optimization of fermentation parameters

Mohammad Karim Khaleghi[1], Iman Shahidi Pour Savizi[1], Nathan Lewis[2], and Seyed Abbas Shojaosadati[1]

[1]Tarbiat Modares University
[2]University of California, San Diego

May 12, 2021

Abstract

Recent noteworthy advances in the development of high-performing microbial and mammalian strains have enabled the sustainable production of bio-economically valuable substances such as bio-compounds, biofuels, and biopharmaceuticals. However, to obtain an industrially viable mass-production scheme, much time and effort are required. The robust and rational design of fermentation processes requires analysis and optimization of different extracellular conditions and medium components, which have a massive effect on growth and productivity. In this regard, knowledge- and data-driven modeling methods have received much attention. Constraint-based modeling (CBM) is a knowledge-driven mathematical approach that has been widely used in fermentation analysis and optimization due to its capabilities of predicting the cellular phenotype from genotype through high-throughput means. On the other hand, machine learning (ML) is a data-driven statistical method that identifies the data patterns within sophisticated biological systems and processes, where there is inadequate knowledge to represent underlying mechanisms. Furthermore, ML models are becoming a viable complement to constraint-based models in a reciprocal manner when one is used as a pre-step of another. As a result, more predictable models are produced. This review highlights the applications of CBM and ML independently and the combination of these two approaches for analyzing and optimizing fermentation parameters.

1. Introduction

Fermentation technology is enjoying a significant moment, due to the potential of metabolic engineering, systems biology, and synthetic biology [1]. Various economically important compounds such as different chemicals, fuels, and biopharmaceuticals can be obtained through fermentation processes. With the purpose of commercialization of any fermentation-based product, the amount of obtained product should meet the market demand [2]. Therefore, optimization of the fermentation parameters (e.g., temperature, pH, medium composition, feeding strategies, etc.) is a critical factor that has an important role in bioprocess overall yield and productivity. Furthermore, fermentation optimization can reduce the overall cost of bioprocess through its impact on downstream processes and purification [3].

Various strategies have been implemented to find the optimal values of fermentation parameters so far. Modeling has always been one of the most popular methods due to its ability to replace expensive laboratory experiments, or at least diminish the amount of them. In this approach, according to the specified algorithms, the output is calculated as a function of given inputs [4]. For instance, the inputs can be media composition, temperature, pH, etc., then the appropriate values of these parameters are optimized to make the desired output. Generally, three types of models are implemented to the problems: purely mechanistic/knowledge-driven, merely data-driven, or a combination of the two [5]. Each of these approaches has its own advantages and disadvantages. For example, data-driven models are black-box models, which do not provide adequate

information on the underlying mechanism [6]. Nevertheless, large datasets may be not incorporated into a model framework smoothly. On the other hand, hypothesis/mechanistic-driven approaches use basic knowledge to extract deeper information from datasets and provide valuable information on the underlying mechanism. Nonetheless, the construction of these models is challenging due to the rapid growth of data. However, in order to construct the most powerful model, it is crucial for the researcher to understand the strengths and weaknesses of these approaches. Moreover, the hybridization of these two approaches might be the most powerful model [7].

Fermentation parameters have a significant effect on cellular metabolism, thus productivity. So, mechanistic analysis of the interaction between environmental conditions and metabolic pathways leads us to fine-tune fermentation parameters in a comprehensive way [8]. There are several mechanistic models for simulating metabolism in the field of systems biology [9]. Among them, constraint-based modeling (CBM) of metabolism is one of the most common approaches [10]. These models are built from a genome-scale metabolic network reconstruction to predict metabolic flux values through optimization techniques such as flux balance analysis (FBA) [11, 12]. To date, genome-scale metabolic models (GEMs) for diverse eukaryotic and prokaryotic organisms and cells have been reconstructed [13] and applied in biotechnology and human health [14, 15]. Such models have been extensively utilized for qualitative mapping of cellular metabolism, predicting metabolic functions, and guiding metabolic engineering designs and bioprocess optimizations toward the desired phenotype [16].

In parallel, machine learning (ML) is a purely data-driven approach with the creation and evolution of algorithms that identify patterns and makes hypothesis or models based on learning from existing data [17, 18]. Because of the rapid increase in omics datasets, many researchers prefer to use machine learning independently to interpret systems biology and metabolic engineering datasets. For instance, genome annotation, host strain selection, pathway discovery, metabolic pathway reconstruction, metabolic flux optimization, multi-omic data integration, and protein modeling can be obtained through machine learning methods [3, 19]. Besides, due to the availability of the large amounts of fermentation parameter values from empirical studies, machine learning algorithms can be implemented directly to this multivariate system to fine-tune the fermentation conditions [20, 21].

Although the applications of each of the two methods separately are constantly increasing, the unique capabilities of each have led to the integration of models with more prediction power and accuracy. Recently, comprehensive reviews of the integration of machine learning algorithms and mechanistic models have been published, indicating a promising outlook for this field of knowledge [22-26]. However, it is worthwhile to review the capabilities of these two methods individually or in combination for fermentation parameter optimization. The basic idea here is that machine learning is a powerful computational tool for analyzing omics data individually or inferring multi-omic relationships. Moreover, as a result of CBM, an additional layer of omics data called fluxomics is created, which can be analyzed by machine learning methods separately or by integrating with other omics data [26].

In the present review, first, we highlight the latest efforts in the literature that utilize CBM as a mechanistic approach for fermentation optimization. Next, we introduce ML as a data-driven method and highlight its recent applications in tuning the fermentation parameters. Finally, we present the studies in which CBM and ML combined to improve the model accuracy for analyzing fermentation conditions.

2. Constraint-based modeling: A mechanistic-driven approach

Cell phenotype depends on various interlaced mechanisms such as metabolism and transcriptional regulation. Kinetic and constraint-based modeling are two main mechanistic approaches in analyzing the principles governing an organism's metabolism and growth [27]. Kinetic models help to understand the dynamic behavior of biological systems. In this approach, the relationship between metabolites is expressed through kinetic laws represented as the ordinary differential equations (ODE) [28]. However, kinetic modeling requires costly and time-consuming efforts to determine sophisticated kinetic parameters (e.g., enzymatic constants and metabolite concentrations). Therefore, the application of this method is limited to small-scale metabolic

models for only extremely well-studied organisms [29]. However, the advances on this method are increasing and significant efforts have been made to build genome-scale kinetic models [30].

Constraint-based modeling (CBM) can overcome the limitations of the kinetic models by reducing the need for complex kinetic parameters. Therefore, this approach has been extensively used for understanding the behavior of genome-wide systems [31]. The main goal of CBM is to build models with high prediction accuracy to analyze the genome-scale networks and shed light on relationships between genotype, phenotype, and environmental conditions [32, 33]. In this section, we first summarize the main concepts of CBM and next, we present recent applications of this method in optimizing the fermentation processes.

2.1 An overview of CBM main concepts

Biological systems such as cellular metabolism are constrained by physiochemical laws, genetics, and the extracellular environment [34]. The most fundamental constraints of metabolism are the mass balance equations for each intracellular metabolite generated from biochemical reaction stoichiometry. Genome-scale network reconstructions are created from all known metabolic reactions within the system of interest. Furthermore, they are improved by additional information, such as gene-protein-reaction (GPR) associations [35]. A valuable manual protocol describes how to generate a high-quality genome-scale metabolic reconstruction from genome sequencing data and how to curate the model with empirical information [12]. A genome-scale network reconstruction can be transformed into a mathematical format. The mathematical representation of such reconstructed networks and implementing further details such as GPR associations is called the genome-scale metabolic model (GEM). This enables the quantitative and qualitative analysis of the GEMs via computational approaches such as constraint-based modeling [36].

Metabolic flux analysis (MFA) and Flux balance analysis (FBA) are two main CBM methods that aim to determine the reaction fluxes (fluxomics) within the metabolic network (**Figure 1**). These methods use a stoichiometric matrix (S) with the size of m * n to calculate the metabolic flux distribution. In the S matrix, each row represents a metabolite (m), and each column represents a metabolic reaction (n). Therefore, under the steady-state condition, the mass balance equation will be as follows: S. v=0. The v vector contains metabolic fluxes, some of which are known and some unknown. MFA is a data-driven method that determines reaction fluxes through experimental measurements. While MFA is useful in small-scale networks, FBA is a beneficial tool for analyzing large-scale networks such as the genome-scale metabolic network [37]. FBA is an optimization method that searches a solution space and maximizes one or more objective functions such as maximum growth rate and metabolite production via a linear programming approach [11]. FBA calculates the single optimal flux distribution or multiple optimal flux distributions in the GEM, which represents the 'state' of the metabolic network that relates to the physiological function generated from the network [38]. However, mass balance constraints alone cannot constitute a unique solution space. Therefore, multiple optimal solutions (i.e., flux vectors) to the problem are obtained. So, additional constraints such as flux capacity, thermodynamic feasibility, gene expression, etc., are imposed to shrink the solution space [39]. Moreover, MFA can combine with FBA to determine internal metabolic fluxes to increase the prediction power [40]. Besides, other forms of FBA and MFA, such as dynamic FBA and MFA, can be used based on the aim of the research [41, 42]. In addition to FBA and MFA, other CBM approaches can be used to rational strain designs and increase product yield. These FBA-based methods aim to determine gene deletion/addition targets, up/down regulations, data integration, and suggest appropriate strategies to increase productivity [43]. These computational methods also can be used in fermentation optimization. For example, up- and down-regulation targets have been used to identify enzyme activators and inhibitors for enhancing the production bound in a regulatory-defined medium (RDM) [44]. COBRA toolbox in MATLAB and COBRApy in python are two platforms for implementing FBA and other related algorithms to GEMs [45, 46].

Another efficient approach to increase the predictive power of the CBM models is the integration of omics data with GEMs. Omics data can be used both to narrow the solution space in the FBA and as a tool to evaluate and validate the model prediction [6]. As a result of integrating omics data with GEMs, context-specific models are created that provide the basis for studying metabolism under different conditions [47, 48].

Assuming that the system is steady-state, substrate concentrations, time, and various kinetic parameters are not taken into account to calculate the metabolic fluxes. Therefore, the predictive accuracy of the CBMs is less than the calculated fluxes resulting from solving ODEs in the kinetic models. As the solution space becomes tighter, the FBA solutions approach the kinetic model solution. Thus, the integration of omics data can overcome the limitations of the CBM over the kinetic models. However, data integration remains a major challenge, and existing methods do not perform at the expected level [49].

2.2 Instances of CBM applications in fermentation optimization

The culture medium is one of the most significant factors affecting cell growth and productivity in bio-based products [4, 50, 51]. Optimization of culture medium and development of proper feeding strategies has always been a critical factor in bioprocess development [2, 52]. Since medium components, nutritional supplements, and culture additives directly affect cellular metabolism [53], CBM can be a suitable approach to analyze the effect of medium components on cell growth and productivity. CBM-based methods have been extensively used to analyze and optimize the culture medium and develop appropriate feeding strategies to overproduce a wide range of products such as recombinant proteins [54-58], biofuels [59-61], lipids [62, 63], chemicals [64-66], and foods [36]. For a more detailed example, Swayambhu et al. used FBA and its derivatives to identify gene deletion/overexpression targets and culture medium components to enhance the production of siderophore compounds in recombinant *Escherichia coli* . First, they used minimization of metabolic adjustment (MoMA) and OptForce algorithms to identify the gene deletion and overexpression targets, respectively. Then, they combined the FBA with Plackett Burman (PB) in silico to identify the amino acids and carbon sources that had a major impact on productivity. Also, it was observed that by changing the rate of nutrient uptake, the priorities of some candidates for gene deletion or overexpression have changed [67]. Fouladiha et al. used a genome-scale metabolic model of the Chinese hamster ovary (CHO) to design feeding strategies to increase the production of a monoclonal antibody. In this study, CBM has been used as pre-step of the design of experiment (DoE) methods to decrease the number of variables and experiments. First, they set the objective function to maximize antibody production and used flux variability scanning based on enforced objective flux (FVSEOF) algorithm to detect reactions whose boundaries had changed. Fifteen exchange reactions were identified related to various media supplements, including three vitamins, seven amino acids, and five other metabolites. PB was then applied to screen for these 15 metabolites, of which threonine and arachidonic acid were identified as the most effective supplements for the culture medium. A more than two-fold increase in protein production was observed through this strategy [68]. In another study, Sarkandy et al. developed a stoichiometric model to predict the most efficient amino acids for enhancing the production of interleukin-2 in fed-batch culture of recombinant *Escherichia coli* . The results showed that the mixture of leucine, aspartic acid, and glycine improves protein productivity by almost two-fold [69]. Recently, Shahidi et al. have comprehensively reviewed the applications of systems biology approaches, including CBM, in chemically defined media formulations for the overproduction of recombinant proteins [70]. FBA has also been used to investigate the effect of glucose, glycerol, and glucose-glycerol dual mixtures on the internal carbon flux distribution in simultaneous ethanol and butanol production [71].

Fermentation conditions also can directly affect the cellular metabolism toward the growth and productivity. Therefore, CBM methods are a convenient and low-cost approach to optimize fermentation conditions. For example, Calic et al. used the MFA method to investigate the effect of pH on the intracellular metabolic network of *Bacillus licheniformis* , a β-lactamase-producing bacterium. In this study, the values of metabolic fluxes related to cell growth, by-products, and desired product production at pH = 6.5, 7, 7.5 were studied. The results show that in the period of cell growth, the by-product fluxes have the highest value at pH = 7 and the lowest value at pH = 7.5, while the change in pH in this period does not have a significant effect on the production of β-lactamase. On the other hand, it has been observed that in the stationary phase and the product formation period, the flux value of the desired product is maximum, and the flux values of by-products are minimum at pH = 7. Finally, this article proposes a pH operation strategy to improve β-lactamase production as follows: First, the initial pH should be set to pH = 7.5 and allowed to decrease to pH = 7 during fermentation, and then kept constant at this value [8]. In another study, Ivarsson et al. manipulated lactate consumption and production by pH alteration during mammalian cell

cultivation. They used FBA to see how pH changes affect lactate metabolism. The results show that by lowering the pH from the standard pH = 7.2 to pH = 6.8, lactate consumption increases, the cell becomes more energy-efficient, and antibody production increases. Moreover, they found that gluconeogenic enzymes regulated the TCA cycle at undesirable pH levels [72]. FBA method constrained by experimentally measured extracellular fluxes has been developed by Sou et al. to investigate the effect of mild hypothermia conditions on antibody glycosylation in the late exponential of a CHO cell. Flux distribution values showed that during the stationary phase at 32 °C, more energy and metabolites are expended to increase cell productivity, which limits most of the resources necessary for antibody glycosylation [73]. The overall overview of CBM applications for fermentation parameters highlighted in this review is presented in **Table 1** .

3. Machine learning: A data-driven approach

Fermentation is a multivariate system in which any number of involved parameters can influence the process outcome [24]. As outlined in the previous section, mechanistic models (e.g., CBM in this review) can lead to fine-tuning some fermentation parameters such as medium composition. Nevertheless, we cannot investigate the effect of all fermentation parameters on productivity through mechanistic approaches. On the other hand, strictly experimental trial-and-error methods are time-intensive and commonly high-priced. Despite the difficulties of such traditional techniques, the large amount of data generated from worthwhile previously fermentation studies provide an appropriate space for data-driven modeling approaches to find the optimal sets of fermentation parameters. Moreover, a rational analysis of large and complex datasets generated from experiments, measurements, and simulations can significantly contribute to an in-depth understanding of the system of interest [74].

Machine learning (ML) is a data-driven approach that uses statistics and probability science to analyze a dataset and discover the hidden relationships between existing data to justify a phenomenon and build a predictive model based on the patterns it learned. In the past, researchers did not distinguish between ML and artificial intelligence (AI), but nowadays, ML is recognized as a subfield of AI [75]. Actually, AI is the industry of developing tools and techniques for ML, while ML uses these tools in various fields such as engineering and science [76]. In the ML process, a problem is first defined on a dataset. Then, a set of preprocessing operations is performed on the dataset based on the defined problem. In the next step, the ML model is created by a user-defined estimator. Finally, the model is validated and evaluated by standard techniques. **Figure 2** shows the general scheme of the machine learning workflow.

3.1 ML process

3.1.1 Dataset preparation

The first step in any machine learning process is to define the desired dataset. These datasets are sometimes called big data [77]. Volume, velocity, variety, veracity, variability, value, and visualization are seven common traits of big data. Therefore, contrary to popular belief, big data is any dataset with at least one of these seven traits, and high-volume datasets are not always required for ML models [78]. However, the greater the amount of appropriate data provided, the more accurate the model built [79]. Each dataset consists of rows and columns, which rows are called samples, and each column can represent a feature or a target value. Features are also called dimensions.**Figure 3** shows an overview of a dataset. Each dataset may contain unknown or outlier values for a variety of reasons, such as high-volume data. Therefore, to prevent modeling errors, it is necessary to perform a primary dataset inspection and identify unusual or missing values. Various machine learning algorithms for missing data imputation and outlier detection have been proposed so far [80]. Input data can be divided into labeled or unlabeled data. While there are one or more target values in labeled data, unlabeled data have no target values. For example, in a labeled dataset, fermentation parameters are features, and productivity is the target value. In these datasets, usually, the aim is to find the relationship between the data and make predictions on new data [77]. In unlabeled datasets, the goal is finding hidden relationships, clustering, or detecting outliers.

3.1.2 Preprocessing

The preprocessing operation must be performed based on the defined problem on the dataset. In this step, using statistical methods, the input data is prepared in such a form that it is desirable for the estimator in the modeling stage [81]. Data transformation is one of the most common preprocessing operations. Various tools and techniques can be implemented for data transformation based on the size, complexity, and structure of the dataset. For instance, data standardization is a transformation method in which the data of each column are transmitted in a specified range (usually between zero and one) [82]. Dimensionality reduction is another beneficial preprocessing method in which data is transferred from a high-dimension space to a low-dimension one preserving some key properties of the original dataset. Principal component analysis (PCA) is the most widely used dimensionality reduction approach that increases data interpretability by constructing new uncorrelated variables summarizing maximize variance [83]. Another prominent preprocessing operation is feature selection which tremendously affects the performance of the ML model. Feature selection is a method that manually or automatically selects only a subset of features that are aligned with the target value in the problem [84].

3.1.3 Modeling

By preparing data in an appropriate format, one can create a model for analyzing and making accurate predictions. First, the original dataset should be divided into training and test sets by specific methods. In the modeling step, the training dataset enters an algorithm, and the algorithm uses statistical and mathematical tools, which are called estimators, to learn and develop predictions. In ML, the process of creating the desired model is called training. Usually, the most challenging part of ML is choosing the right estimator based on different types of data and problems. The scikit-learn library provides a broad range of estimators along with a procedure on how to select the best estimator among them (www.scikit-learn.org). Some important machine learning estimators including linear regression, k-nearest neighbors, support vector machines (SVMs), decision trees, random forest, Gaussian process (GP), fuzzy logic, and artificial neural networks (ANNs) are defined in **Table 2** . Moreover, ensemble methods aim to fuse the prediction of several single estimators to improve the precision and accuracy of the model [85]. For example, the random forest is an ensemble modeling method in which several decision trees are used to predict the outcome [86]. It is worth noting that each estimator has its tunable parameters depending on the type of data and the problem.

3.1.4 Validation and evaluation

Model evaluation is first conducted on the training dataset to measure how well the model can predict previously unseen data [87]. For instance, k-fold cross-validation is a standard method that evaluates the estimator's performance by randomly splitting the training dataset into training and test sets. For better understanding, in this method, the training dataset is divided randomly into k equal parts, called k folds. Then, the model runs k times, and in each round, one particular fold is used as the test set and k-1 folds as training ones [88]. The accuracy of the model is computed for each test fold. When k-fold cross-validation is operated, one can see how sensitive the model is to the training dataset. Next, by evaluating the accuracy, the model's parameters can be tuned to improve its generalization performance for unseen data [89, 90]. The final evaluation of the model is conducted with test data, and the predictive ability of the model is quantified by different evaluation metrics such as accuracy, sensitivity, specificity, precision, and recall [91, 92]. **Figure 4** shows how evaluation methods can help to improve the model's performance.

3.2 ML categories

ML methods are broadly divided into supervised, unsupervised, and semi-supervised learning [93]. Unsupervised learning aims to uncover the hidden patterns and deduce the structures of unlabeled training data [18]. Unsupervised learning approaches cluster subgroups of data with similar properties or features into separate categories. Moreover, dimensionality reduction methods mentioned earlier as preprocessing operations are unsupervised learning methods that reduce the number of old features and create new principal features with minimum information loss [94]. However, because there are no target values in unsupervised learning methods, they cannot build an independent predictive model [95]. On the other hand, supervised learning is applied to learn and discover associations relationships between features and target values in a labeled

dataset [96]. Therefore, the built model can be used as a predictive one to test previously unseen data. Two main supervised methods are classification for discrete class labels and regression for numerical quantities [97]. It is noteworthy that the output of unsupervised learning methods (e.g., low-dimension space features) can be used as the input of supervised learning algorithms. For semi-supervised learning, there are both labeled and unlabeled data in the training set; But usually, the amount of unlabeled data is more. In this method, one can use labeled data to create labels for unlabeled ones [98].

3.3 Applications of ML algorithms for fermentation analysis and optimization

ANNs have been used successfully in several studies in the field of fermentation prediction and optimization (**Table 3**). The predictive capacity comparison of ANN and RSM has been studied by Nelofer et al. for the lipase production process by a recombinant*Escherichia coli* [99]. In this study, fermentation parameters were optimized based on experimental lipase production data. As a result, ANN showed better performance over RSM for both R2 and adjusted-R2 values. Moreover, absolute average deviation (AAD) and root mean square error (RMSE) in the ANN model gave lower values, indicating the high accuracy of ANN. Instead of comparing ANN and RSM, integration of these two strategies has also attracted much attention in recent years. For instance, in a recent study, Wang et al. proposed an ANN-RSM methodology to overcome the pure RSM failure in predicting complex nonlinear systems. They used original experimental datasets to train and validate an ANN model and produce response surface models to analyze the effect of critical parameters in dark hydrogen fermentation. The constructed model showed good and reliable results for this nonlinear and noisy process [100]. Genetic algorithm (GA) is a global search optimization method inspired by natural selection theory. GA usually has been coupled with ANN to find the optimum values of fermentation parameters used in model training. Recently, Unni et al. employed ANN together with GA to optimize medium composition for the production of human interferon-gamma (hIFN-γ) using a recombinant *Kluyveromyces lactis* [101]. Recently, an on-line μ-stat strategy was proposed for controlling methanol feeding in a fed-batch process of Recombinant*Pichia pastoris* . In this study, Tavasoli et al. employed MLP3 neural network (a class of ANNs) to reconstruct and adjust the controller's performance. Consequently, a significant enhancement was observed in the production of human recombinant alpha 1-antitrypsin (A1AT) [102].

SVMs are another popular method for training experimental fermentation data and predicting the process outcome. One specific advantage that SVMs have over ANNs is that they always find the global optimum solution, while ANNs may fall into the local optimum. Moreover, SVMs are effective for problems with a small number of samples. The predictive capabilities of SVM and ANN were compared recently by Zhang et al. In this investigation, SVM and ANN were used to build models for predicting biomass yield, lipid production, and COD removal rate in a microbial lipid fermentation. The results demonstrated that the SVM linked with the genetic algorithm performed better over ANN with a small number of samples [103]. In another study, the least-square SVM (LS-SVM), a modified SVM, was coupled with orthogonal experimental design (OED) to map the relationship between process parameters as inputs and cumulative biogas production (CBP) as the output for corn stalks anaerobic fermentation with only nine samples. In this study, the LS-SVM parameters were optimized by the grid search method. The results showed that using this optimization method as an alternative to pure OED increases CBP by 14.13% [104].

Other ML methods also have shown reliable results in fermentation prediction and optimization. Kennedy et al. investigated the capabilities of fuzzy logic as a tool for media formulation. They found that this method can save 63% of the experiments and the remaining experiments are adequate for media design. They found that the selection of correct number of fuzzy logic rules is critical for enhancing model accuracy [105]. In another study, Melcher et al. utilized random forest and ANN for biomass and recombinant protein modeling in a fed-batch *Escherichia coli* process. Online fermentation parameters and two-dimensional (2D) fluorescence spectroscopy were used for dry cell mass and productivity prediction. The hybrid model accuracy reached about $\pm$4% for dry cell mass and $\pm$12% for protein concentration [106]. Masampally et al. employed Gaussian process regression (GPR) in fed-batch fermentation of yeast *saccharomyces cerevisiae* to predict biomass concentration. In this study, three cascade sub-models were developed to predict gas hold-up, dissolved oxygen (DO), and biomass storage, respectively. Validation experiments were eventually perfor-

med [107]. Recently, using the k-nearest-neighbor (KNN) method, a 1.64-fold improvement in *Penicillium brevicompactum*fermentation producing mycophenolic acid (MPA) has been obtained [108].

4. Incorporation of constraint-based modeling and machine learning

In recent years, the advances in high-throughput devices and the rapid growth of omics data provide a unique opportunity to depict biological samples at multiple layers. Omics data generated from high-throughput technologies are big data that can be analyzed individually or inferred as multi-omic relationships through ML algorithms to gain more biological insight [109-111]. Common omics datasets include genomics, transcriptomics, proteomics, and metabolomics produced from DNA sequencing, RNA sequencing and microarrays, and mass spectrometry, respectively [112]. Furthermore, fluxomics is an additional layer of omics generated from CBM approaches and includes metabolic flux distribution values, thus representing the metabolic phenotype [113]. Consequently, the integration of ML (for omics and multi-omics analysis) and CBM (for generating fluxomics) looks promising for analyzing a biological system such as cellular metabolism. However, the capabilities of this hybrid approach are just now being explored, and more research in this area is needed. In this section, first, we briefly summarize the ways that ML and CBM can be coupled. Next, the most prominent studies aiming at analysis and optimization of fermentation parameters by ML-CBM approaches are reviewed.

4.1 Approaches for CBM and ML integration

The integration of ML and CBM can be conducted in three chief ways [26]: (a) The output of CBM is the flux distribution that indicates the metabolic state of the cell. This fluxomic data can be trained directly through ML methods to obtain more biological insight into the desired system (**Figure 5A**). (b) ML is an effective tool for merging and analyzing heterogeneous omics datasets beyond ML applications to single omics. By combining these multi-omics datasets with GEMs, context-specific models are generated. More accurate flux values obtained from context-specific GEMs can be re-integrated with experimental omics data for further predictions (**Figure 5B**). (c) CBM models and fluxomic data can be produced directly by introducing omics or multi-omics datasets into ML algorithms. All of the three mentioned methods might be operated by supervised or unsupervised algorithms (**Figure 5C**).

4.2 Instances of CBM coupled with ML for fermentation analysis and optimization

Routinely, CBM uses genetic and environmental conditions as inputs to predict metabolic flux distributions. However, Sridhara et al. investigated whether they could infer bacterial growth conditions from internal fluxomics in an inverse manner. For this reason, the prediction conducted using a simple linear regression. The results showed that using the intracellular flux values, carbon and nitrogen sources utilized in the initial culture medium could be predicted even with a small number of impurities [114]. In a recent study, Oyetunde et al. extracted over 1,200 curated bioprocess datasets from 100 articles to predict microbial factories' performance (yield, titer, and rate). The authors generated additional flux-based features from a CBM model to augment ML input data. Next, they applied ensemble methods to alleviate data challenges such as sparse, non-standardized, and incomplete datasets. The developed ML-CBM model could predict an engineered*Escherichia coli* performance with high accuracy [115]. In 2016, Wu et al. developed MFlux, an online platform, for predicting bacterial central metabolism. The authors used ML approaches (SVM, KNN, and decision tree) to train previously experimental data, including substrate types, bioprocess strategies, and genetic modifications from about 100 13C-MFA articles. MFlux outputs can be used as inputs for FBA to reduce the solution space, thus improving the model's accuracy [116].

Most recently, a novel CBM-ML hybridization approach for time-course controlling nutrients availability in a fed-batch CHO cell culture has been developed. For this reason, Schinn et al. used ML as a tool to overcome CBM limitations, such as optimal metabolism considerations and steady-state assumptions. In this study, cell density, product titer, glucose, lactate, glutamine, and glutamate concentrations were used as constraints for the FBA solution. The metabolic model calculated the initial consumption rates of proteogenic amino acids. Next, a series of linear regressions were used to refine the predictions. Finally, using a sigmoid function, the refined consumption rates were fit to a time-course dependent profile. The model was able to

correctly forecast the concentrations of 13 out of 18 amino acids [117].

Essential genes are genes that are critical for cell viability and growth. Gene essentiality is not an intrinsic trait of a gene. But instead, it can be influenced by environmental and genetic contexts [118]. Nandi et al. developed an SVM-based model named SVM-RFE to classify *Escherichia coli* genes as essential or non-essential. The model input included a mixture of genotypic and phenotypic features, i.e., gene and protein sequences, topological network, and gene expression. Then, they employed flux coupling analysis (FCA) to generate flux-based features to consider gene adaptability in different environmental conditions. SVM-RFE was trained on 4094 reaction-gene combinations with 64 features. The model could successfully capture the minimal set of essential genes in various environmental conditions with high accuracy [119]. This study shows the importance of selecting and describing appropriate features in an ML study.

In the context of multi-omics integration, Zampieri et al. employed a combination of CBM and ML to predict lactate production, a secondary metabolite, in CHO cell culture. In this study, transcriptomics data from different culture conditions were integrated with fluxomics data from in-silico genome-scale modeling to construct a data-driven framework. The results showed an improving performance over the predictive power of pure transcriptomic analysis [120]. Similarly, Vijayakumar et al. proposed a machine learning pipeline integrated with genome-scale modeling to improve phenotypic prediction in a lipid-producing cyanobacterium. First, they extracted RNA sequencing data from 23 different growth conditions to develop condition-specific GEMs via transcriptomics data integration. Then, FBA was performed to obtain context-specific fluxomic data. The preprocessing stage was conducted to incorporate fluxomics into experimental transcriptomics data. PCA, k-mean clustering, and LASSO regression were used to identify the dataset's key features. As a result, a data-driven multi-view model was developed with a high phenotype predictive accuracy [121]. This strategy also has been adapted to predict yeast S. cerevisiae growth rate. In this study, fluxomics, generated from parsimonious flux balance analysis (pFBA), were coupled with transcriptomics to train neural networks [122].

5. Challenges and perspectives

It is now clear that fermentation optimization and control are needed to achieve more efficient and economic bioprocess. Therefore, different holistic strategies have been developed to obtain specifically tailored fermentation parameters. In this review, we briefly summarized the applications of two important modeling methods for analyzing and optimizing fermentation parameters: (a) Constraint-based modeling (CBM) and (b) machine learning (ML). The former is a mechanistic method that aims to get more insight into the biological system of metabolism. The latter is a data-driven approach in which a specified algorithm can learn from previous data and make predictions with minimal human intervention. ML can be used to analyze fermentation parameters directly or to infer high-dimensional omics datasets. Furthermore, recent investigations showed that ML can be integrated with CBM to improve predictive power and get more biological insights. However, despite the advances in CBM, ML, and CBM-ML applications in fermentation analysis and optimization, several limitations remain to be discussed.

Here we review some of the major challenges in developing CBM and ML models. CBM can only estimate the metabolic flux distribution relying on an optimal steady-state assumption in which it operates well only in ideal limited time scales. Moreover, it does not account for metabolite concentrations, enzyme kinetics, and regulations. Therefore, CBM cannot always predict accurate metabolic fluxes [11]. Kinetic models of metabolism that can overcome these shortcomings have been successfully developed to analyze the effect of genetic and environmental factors at genome scales [30]. However, difficult and computationally expensive methods have limited kinetic model performance in an intracellular environment [123]. An alternative approach to kinetic models is using omics and multi-omics datasets to generate context-specific metabolic models and derive meaningful insights [49, 124]. To this end, several algorithms have been developed with different assumptions and predictive capabilities [125, 126]. Nevertheless, omics datasets also have some limitations, such as heterogeneity of individual omics, the necessity of intensive analysis, differences in representation formats [124], lack of mechanistic knowledge, and inefficient genome-scale integration tools. Thus, the construction of context-specific GEMs remains challenging. Therefore, many researchers prefer to

build data-driven ML models in order to analyze metabolic networks [7]. Moreover, thanks to the significant amount of fermentation studies and advances in measuring tools, ML methods have been widely used to optimize fermentation media and conditions. Nevertheless, ML models are black-box models, which use previously experimental datasets and do not provide sufficient information on the underlying mechanism [6].

As described above, CBM and ML both can be used as powerful tools for analyzing metabolic networks and fermentation parameters. However, the remarkable capabilities of each method have shown promise for the construction of combined CBM-ML methods. The collaboration between ML and CBM is a reciprocal process. In other words, ML can be applied to the CBM input datasets and increase the predictive power of the metabolic model. Conversely, CBM is a practical tool for the generation of a new layer of omics data called fluxomics to improve the interpretability of a data-driven ML model [23]. However, to take full advantage of both CBM and ML, several challenges needed to be addressed. First, CBM-derived fluxomic data require several preprocessing steps to integrate with multi-omic data to obtain suitable biological data. This target is restricted due to the heterogeneous and high-dimensional datasets [111]. Second, despite the advances in genome-scale metabolic reconstructions, appropriate high-throughput data are only available for a small group of microorganisms [19]. Third, the results of the integrated models, although very accurate, are not necessarily appropriate for large-scale industrial fermentations. Finally, the limitations of each ML and CBM method are still not fully addressed in the integrated models.

Recent studies represent a hopeful future for improving fermentation processes and obtaining valuable biological products through the methods reviewed in this article. In our view, to overcome the challenges, it is required to enhance our biological knowledge in parallel with the development of novel mathematical and computational tools. This is possible through collaboration between chemical engineers, biologists, physicists, and computer engineers. For example, the integration of high-quality metabolic regulatory networks can improve the prediction power of CBM models. Moreover, further developments in high-throughput techniques and more effective methods are needed to integrate omics data with GEMs.

Acknowledgments

This work was supported by Tarbiat Modares University (Grant No. IG-39702); and generous support from the Novo Nordisk Foundation (NNF20SA0066621).

Conflict of interest

The authors declare no financial or commercial conflict of interest.

References :

[1] Choi, K. R., Jang, W. D., Yang, D., Cho, J. S.*, et al.* ,*Trends Biotechnol* 2019, *37* , 817.

[2] Singh, V., Haque, S., Niwas, R., Srivastava, A.*, et al.* ,*Front Microbiol* 2017, *7* , 2087.

[3] Kim, G. B., Kim, W. J., Kim, H. U., Lee, S. Y., *Curr Opin Biotechnol* 2020, *64* , 1.

[4] Galbraith, S. C., Bhatia, H., Liu, H., Yoon, S., *Curr Opin Chem Eng* 2018, *22* , 42.

[5] Shih, W., Chai, S., *Academy of Management Proceedings* , Academy of Management Briarcliff Manor, NY 10510 2016, p. 14843.

[6] Hyduke, D. R., Lewis, N. E., Palsson, B. O., *Mol Biosyst*2013, *9* , 167.

[7] Presnell, K. V., Alper, H. S., *Biotechnol J* 2019,*14* , 1800416.

[8] Calik, P., Ileri, N., *Chem Eng Sci* 2007, *62* , 5206.

[9] Nielsen, J., *Annu Rev Biochem* 2017, *86* , 245.

[10] Bordbar, A., Monk, J. M., King, Z. A., Palsson, B. O.,*Nat Rev Genet* 2014, pp. 107.

[11] Orth, J. D., Thiele, I., Palsson, B. O., *Nat Biotechnol*2010, *28* , 245.

[12] Thiele, I., Palsson, B. O., *Nat Protoc* 2010, *5* , 93.

[13] Reed, J. L., *The Chemistry of Microbiomes: Proceedings of a Seminar Series* , National Academies Press (US) 2017.

[14] Chowdhury, S., Fong, S. S., *Curr Opin Biotechnol* 2020, pp. 267.

[15] Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., Lee, S. Y.,*Genome Biol* 2019, *20* , 121.

[16] Zhang, J., Petersen, S. D., Radivojevic, T., Ramirez, A.*, et al.* , *Nat Commun* 2020, *11* , 4880.

[17] Rana, P., Berry, C., Ghosh, P., Fong, S. S., *Curr Opin Biotechnol* 2020, *64* , 85.

[18] Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., Drăghici, S., *PLoS Comput Biol* 2007, *3* , e116.

[19] Helmy, M., Smith, D., Selvarajoo, K., *Metab Eng Commun*2020, *11* , e00149.

[20] Chen, F., Li, H., Xu, Z., Hou, S., Yang, D., *Electron J Biotechnol* 2015, *18* , 273.

[21] Suryawanshi, N., Sahu, J., Moda, Y., Eswari, J. S., *Prep Biochem Biotechnol* 2020, *50* , 1031.

[22] Antonakoudis, A., Barbosa, R., Kotidis, P., Kontoravdi, C.,*Comput Struct Biotechnol J* 2020.

[23] Kim, Y., Kim, G. B., Lee, S. Y., *Curr Opin Syst Biol*2021, *25* , 42.

[24] Lawson, C. E., Martí, J. M., Radivojevic, T., Jonnalagadda, S. V. R.*, et al.* , *Metab Eng* 2021, *63* , 34.

[25] Oyetunde, T., Bao, F. S., Chen, J.-W., Martin, H. G., Tang, Y. J., *Biotechnol Adv* 2018, *36* , 1308.

[26] Zampieri, G., Vijayakumar, S., Yaneske, E., Angione, C.,*PLoS Comput Biol* 2019, *15* , e1007084.

[27] Yasemi, M., Jolicoeur, M., *Processes* 2021, *9* , 322.

[28] Kim, O. D., Rocha, M., Maia, P., *Front Microbiol* 2018,*9* , 1690.

[29] Traustason, B., Cheeks, M., Dikicioglu, D., *Int J Mol Sci* 2019, *20* , 5464.

[30] Srinivasan, S., Cluett, W. R., Mahadevan, R., *Biotechnol J* 2015, *10* , 1345.

[31] Volkova, S., Matos, M. R. A., Mattanovich, M., Marín de Mas, I., *Metabolites* 2020, *10* , 303.

[32] King, Z. A., Lloyd, C. J., Feist, A. M., Palsson, B. O.,*Curr Opin Biotechnol* 2015, *35* , 23.

[33] Maia, P., Rocha, M., Rocha, I., *Microbiol Mol Biol Rev*2016, *80* , 45.

[34] Li, S., Richelle, A., Lewis, N. E., in: Lee, G. M., Faustrup Kildegaard, H., Lee, S. Y., Nielsen, J., Stephanopoulos, G. (Eds.),*Cell Culture Engineering* , Wiley 2019, pp. 73.

[35] Reed, J. L., Palsson, B. Ø., *Genome Res* 2004, *14* , 1797.

[36] Rau, M. H., Zeidan, A. A., *Biochem Soc Trans* 2018,*46* , 249.

[37] Huang, Z., Lee, D.-Y., Yoon, S., *Biotechnol Bioeng* 2017,*114* , 2717.

[38] O'Brien, E. J., Monk, J. M., Palsson, B. O., *Cell* 2015,*161* , 971.

[39] Reed, J. L., *PLoS Comput Biol* 2012, *8* , e1002662.

[40] Suthers, P. F., Maranas, C. D., *AlChE J* 2020, *66* .

[41] Anand, S., Mukherjee, K., Padmanabhan, P., *Biotechnol Genet Eng Rev* 2020, pp. 1.

[42] Antoniewicz, M. R., *Curr Opin Biotechnol* 2013, pp. 973.

[43] Long, M. R., Ong, W. K., Reed, J. L., *Curr Opin Biotechnol* 2015, *34* , 135.

[44] Motamedian, E., Sarmadi, M., Derakhshan, E., *Process Biochem* 2019, *87* , 10.

[45] Ebrahim, A., Lerman, J. A., Palsson, B. O., Hyduke, D. R.,*BMC Syst Biol* 2013, *7* , 74.

[46] Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N.*, et al.* , *Nat Protoc* 2019, *14* , 639.

[47] Fouladiha, H., Marashi, S.-A., *J Biomed Inform* 2017,*68* , 35.

[48] Rai, A., Saito, K., *Curr Opin Biotechnol* 2016,*37* , 127.

[49] Ramon, C., Gollub, M. G., Stelling, J., *Essays Biochem*2018, *62* , 563.

[50] Reinhart, D., Damjanovic, L., Kaisermayer, C., Kunert, R.,*Appl Microbiol Biotechnol* 2015, *99* , 4645.

[51] Ritacco, F. V., Wu, Y., Khetan, A., *Biotechnol Progr*2018, *34* , 1407.

[52] Maghsoudi, A., Hosseini, S., Shojaosadati, S. A., Vasheghani-Farahani, E.*, et al.* , *Biotechnol Bioprocess Eng*2012, *17* , 76.

[53] Kishishita, S., Nishikawa, T., Shinoda, Y., Nagashima, H.*, et al.* , *J Biosci Bioeng* 2015, *119* , 700.

[54] Calmels, C., McCann, A., Malphettes, L., Andersen, M. R.,*Metab Eng* 2019, *51* , 9.

[55] Emenike, V. N., Schenkendorf, R., Krewer, U., *Comput Chem Eng* 2018, *118* , 1.

[56] Huang, Z., Xu, J., Yongky, A., Morris, C. S.*, et al.* ,*Biochem Eng J* 2020, *160* , 107638.

[57] Irani, Z. A., Maghsoudi, A., Shojaosadati, S. A., Motamedian, E., *Biochem Eng J* 2015, *98* , 1.

[58] Meadows, A. L., Karnik, R., Lam, H., Forestell, S., Snedecor, B., *Metab Eng* 2010, *12* , 150.

[59] Bideaux, C., Montheard, J., Cameleyre, X., Molina-Jouve, C., Alfenore, S., *Appl Microbiol Biotechnol* 2016, *100* , 1489.

[60] Pham, N., Reijnders, M., Suarez-Diez, M., Nijsse, B.*, et al.* , *Biotechnol Biofuels* 2021, *14* , 1.

[61] Zhao, X., Kasbi, M., Chen, J., Peres, S., Jolicoeur, M.,*Biotechnol Bioeng* 2017, *114* , 2907.

[62] Boyle, N. R., Morgan, J. A., *BMC Syst Biol* 2009,*3* , 1.

[63] Parichehreh, R., Gheshlaghi, R., Mahdavi, M. A., Elkamel, A.,*Biochem Eng J* 2019, *141* , 131.

[64] Aminian-Dehkordi, J., Mousavi, S. M., Marashi, S. A., Jafari, A., Mijakovic, I., *Front Bioeng Biotechnol* 2020, *8* , 528.

[65] Fan, S., Zhang, Z., Zou, W., Huang, Z.*, et al.* , *J Biotechnol* 2014, *169* , 15.

[66] Lee, K., Park, J., Kim, T., Yun, H., Lee, S., *Microb Cell Fact* 2010, *9* , 94.

[67] Swayambhu, G., Moscatello, N., Atilla-Gokcumen, G. E., Pfeifer, B. A., *iScience* 2020, *23* , 101016.

[68] Fouladiha, H., Marashi, S.-A., Torkashvand, F., Mahboudi, F.*, et al.* , *Bioprocess Biosystems Eng* 2020, *43* , 1381.

[69] Yegane-Sarkandy, S., Farnoud, A. M., Shojaosadati, S. A., Khalilzadeh, R.*, et al.* , *Biotechnol Appl Biochem* 2009,*54* , 31.

[70] Savizi, I. S. P., Soudi, T., Shojaosadati, S. A., *Appl Microbiol Biotechnol* 2019, *103* , 8315.

[71] Kaushal, M., Chary, K. V. N., Ahlawat, S., Palabhanvi, B.*, et al.* , *Bioresour Technol* 2018, *249* , 767.

[72] Ivarsson, M., Noh, H., Morbidelli, M., Soos, M.,*Biotechnol Progr* 2015, *31* , 347.

[73] Sou, S. N., Sellick, C., Lee, K., Mason, A.*, et al.* ,*Biotechnol Bioeng* 2015, *112* , 1165.

[74] Gupta, M. K., Misra, K., *Netw Model Anal Health Inform Bioinform* 2016, *5* , 4.

[75] Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., De Felice, F., *Sustainability* 2020, *12* , 492.

[76] Venkatasubramanian, V., *AlChE J* 2019, *65* , 466.

[77] Muller, A. C., Guido, S., 2016.

[78] Erl, T., Khattak, W., Buhler, P., Prentice Hall Boston 2016.

[79] Volk, M. J., Lourentzou, I., Mishra, S., Vo, L. T.*, et al.* , *ACS Synth Biol* 2020, *9* , 1514.

[80] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.*, et al.* , *J Mach Learn Res* 2011, *12* , 2825.

[81] Cielen, D., Meysman, A., Ali, M., 2016.

[82] Kumari, B., Swarnkar, T., *Advanced Computing and Intelligent Engineering* , Springer 2020, pp. 309.

[83] Jolliffe, I. T., Cadima, J., *Philos T R Soc A* 2016,*374* , 20150202.

[84] Bolon-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A.,*Progress in Artificial Intelligence* 2016, *5* , 65.

[85] Ardabili, S., Mosavi, A., Varkonyi-Koczy, A. R., *Lecture Notes in Networks and Systems* 2020.

[86] Breiman, L., *Machine Learning* 2001, *45* .

[87] Vabalas, A., Gowen, E., Poliakoff, E., Casson, A. J.,*PLoS One* 2019, *14* , e0224365.

[88] Zhang, D., Del Rio-Chanona, E. A., Petsagkourakis, P., Wagner, J., *Biotechnol Bioeng* 2019, *116* , 2919.

[89] Caglar, M. U., Hockenberry, A. J., Wilke, C. O., *PLoS One* 2018, *13* , e0206634.

[90] Tokuyama, K., Shimodaira, Y., Terawaki, T., Kusunose, Y.*, et al.* , *J Biosci Bioeng* 2020, *130* , 409.

[91] Danjuma, K. J., *arXiv preprint arXiv:1504.04646* 2015.

[92] Saito, T., Rehmsmeier, M., *Bioinformatics* 2017,*33* , 145.

[93] Huang, S., Chaudhary, K., Garmire, L. X., *Front genet*2017, *8* , 84.

[94] Gilpin, W., Huang, Y., Forger, D. B., *Curr Opin Syst Biol* 2020, *22* , 1.

[95] Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., Collins, J. J., *Cell* 2018, pp. 1581.

[96] Smiatek, J., Jung, A., Bluhmki, E., *Trends Biotechnol*2020.

[97] Cuperlovic-Culf, M., *Metabolites* 2018, *8* , 4.

[98] Kim, M., Rai, N., Zorraquino, V., Tagkopoulos, I., *Nat Commun* 2016, *7* , 13090.

[99] Nelofer, R., Ramanan, R. N., Rahman, R. N. Z. R. A., Basri, M., Ariff, A. B., *J Ind Microbiol Biotechnol* 2012, *39* , 243.

[100] Wang, Y., Yang, G., Sage, V., Xu, J.*, et al.* ,*Environ Prog Sustain Energy* 2021, *40* .

[101] Unni, S., Prabhu, A. A., Pandey, R., Hande, R., Veeranki, V. D., *Can J Chem Eng* 2019, *97* , 843.

[102] Tavasoli, T., Arjmand, S., Ranaei Siadat, S. O., Shojaosadati, S. A., Sahebghadam Lotfi, A., *Biochem Eng J* 2019, *144* , 18.

[103] Zhang, L., Chao, B., Zhang, X., *Bioresour Technol* 2020,*301* , 122781.

[104] Dong, C., Chen, J., *Bioresour Technol* 2019, *271* , 174.

[105] Kennedy, M. J., Spooner, N. R., *Biotechnol Tech* 1996,*10* , 47.

[106] Melcher, M., Scharl, T., Spangl, B., Luchner, M.*, et al.* , *Biotechnol J* 2015, *10* , 1770.

[107] Masampally, V. S., Pareek, A., Runkana, V., *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* , IEEE 2018, pp. 128.

[108] Patel, G., Patil, M. D., Tangadpalliwar, S., Nile, S. H.*, et al.* , *Ultrasound Med Biol* 2021, *47* , 777.

[109] Misra, B. B., Langefeld, C., Olivier, M., Cox, L. A., *J Mol Endocrinol* 2019, *62* , R21.

[110] Noor, E., Cherkaoui, S., Sauer, U., *Curr Opin Syst Biol*2019, *15* , 39.

[111] Xu, C., Jackson, S. A., *Genome Biol* 2019, *20* , 76.

[112] Horgan, R. P., Kenny, L. C., *Obstet Gynecol* 2011,*13* , 189.

[113] Kromer, J., Quek, L.-E., Nielsen, L., *Aust Biochem*2009, *40* , 17.

[114] Sridhara, V., Meyer, A. G., Rai, P., Barrick, J. E.*, et al.* , *PLoS One* 2014, *9* , e114608.

[115] Oyetunde, T., Liu, D., Martin, H. G., Tang, Y. J., *PLoS One* 2019, *14* , e0210558.

[116] Wu, S. G., Wang, Y., Jiang, W., Oyetunde, T.*, et al.* ,*PLoS Comput Biol* 2016, *12* , e1004838.

[117] Schinn, S. M., Morrison, C., Wei, W., Zhang, L.*, et al.* , 2021, *118* , 2118.

[118] Larrimore, K. E., Rancati, G., *Curr Opin Genet Dev*2019, *58-59* , 55.

[119] Nandi, S., Subramanian, A., Sarkar, R. R., *Mol Biosyst*2017, *13* , 1584.

[120] Zampieri, G., Coggins, M., Valle, G., Angione, C., *The 2nd International Electronic Conference on Metabolomics* , MDPI 2017, p. 4993.

[121] Vijayakumar, S., Rahman, P. K. S. M., Angione, C.,*iScience* 2020, *23* , 101818.

[122] Culley, C., Vijayakumar, S., Zampieri, G., Angione, C.,*PNAS* 2020, *117* , 18869.

[123] Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., Jirstrand, M., *Metab Eng* 2014, pp. 38.

[124] Subramanian, I., Verma, S., Kumar, S., Jere, A., Anamika, K.,*Bioinform Biol Insights* 2020, *14* , 1177932219899051.

[125] Blazier, A. S., Papin, J. A., *Front Physiol* 2012.

[126] Richelle, A., Chiang, A. W. T., Kuo, C.-C., Lewis, N. E.,*PLoS Comput Biol* 2019, *15* , e1006867.

Table 1. Overview of constraint-based modeling applications for analysis and optimization of fermentation parameters demonstrated in this review.

Parameter	Aim	Achievement	Ref
Culture medium	Combining the FBA with Plackett Burman (PB) for a recombinant *Escherichia coli*	Identifying the key amino acids and carbon sources for successful improvement in siderophore production	[67]
Culture medium	Using the FVSEOF algorithm to design feeding strategies for CHO cells	Suggesting threonine and arachidonate as supplements by the metabolic model and a two-fold increase in antibody production	[68]
Culture medium	Designing a strategy for amino acid supplementation by a stoichiometric model	Enhancing the production of interleukin-2 by adding a mixture of leucine, aspartic acid, and glycine	[69]

Parameter	Aim	Achievement	Ref
Culture medium	Performing FBA for *C. sporogenes* NCIM on glucose, glycerol & their mixture	Glucose and glucose-glycerol mixture improve the growth and the butyric acid production. Glycerol reduces the growth and improves ethanol yield	[71]
pH	MFA used for *Bacillus licheniformis* to study the effect of pH on metabolic flux distribution	Proposing a pH-controlled strategy to improve β-lactamase production	[8]
pH	FBA used to gain mechanistic insight into the question of how pH affects lactate metabolism for a mammalian cell	Lowering the pH increases lactate consumption, antibody production and makes the cell more energy-efficient	[72]
Temperature	FBA implemented to study the effect of mild hypothermia conditions on the CHO cell behaviors	Formation of premature glycosylated antibodies during the stationary phase at 32 °C	[73]

Table 2. A brief introduction to the machine learning estimators (algorithms) that considered in this review.

Estimator (Algorithm)	Task	Description	Relevant studies
Linear regressions	Regression	The simplest form of regression models. Attempts to minimize the mean squared error (MSE) between actual target values and the target values estimated by fitting a linear equation to the training set.	[114, 120, 121, 123]
k-nearest neighbors (KNN)	Classification Regression	A simple ML algorithm that stores k nearest neighbor samples in a dataset (k=1, 2, 3, ...) based on the feature similarities.	[81, 108, 116]

Estimator (Algorithm)	Task	Description	Relevant studies
Support vector machines (SVMs)	Classification Regression	The objective of SVM is to transform each data in an n-dimensional space (n: number of features) and separate data points into two categories in a manner that maximizes the width of the gap between the nearest observations. SVMs are divided into two main groups: Support vector regressors (SVRs) and support vector classifiers (SVCs).	[103, 104, 115, 116, 119]
Decision trees and random forest	Classification Regression	Decision trees provide a tree-shaped structure, including nodes and branches. The algorithm makes decisions based on a hierarchy of if/else questions. Each node classifies the input depending on the question, and the branches are the final decisions representing the output. Random forest is an ensemble method that collects many randomly made decision trees. This algorithm aims to overcome the limitations of each individual decision tree and averages their results.	[106, 115, 116]
Gaussian process (GP)	Classification Regression	GP is a probabilistic non-parametric method that aims to make predictions and provide uncertainty information on the estimations based on Bayes' rule. Gaussian process classification (GPC) and Gaussian process regression (GPR) are two main groups of GP.	[107]

Estimator (Algorithm)	Task	Description	Relevant studies
Fuzzy logic (FL)	Regression	FL resembles the pattern of human reasoning for solving problems considering all available possibilities between Yes and No.	[105]
Artificial neural networks (ANNs)	Classification Regression	ANNs are one of the most popular ML algorithms which are inspired by the human brain information process. ANNs are consist of interconnected neurons arranged in an input layer, a series of hidden layers, and an output layer. In these algorithms, each neuron makes decisions and gives it to other neurons in the next layer. The weighted sum of the inputs is calculated by an activation function. Then, the procedure tends to update the weights in order to minimize the prediction errors.	[99-102, 106, 115, 117, 122]

Table 3. Overview of machine learning applications for analysis and optimization of fermentation parameters demonstrated in this review.

Host	Aim	Input features	Achievement	Ref
Escherichia coliBL21	Comparing RSM and ANN for optimizing lipase production	Glucose, sodium chloride (NaCl), temperature and induction time	ANN showed better R^2 and adjusted-R^2 values, and lower AAD and RMSE values	[99]
Sludge	Integrating RSM and ANN for improving biohydrogen titer	Carbon sources, metal cofactor Fe^0, pH, and sludge concentration	Developing a robust, cost-effective, and reliable optimization methodology	[100]

Host	Aim	Input features	Achievement	Ref
Kluyveromyces lactis	Using ANN coupled with GA for media optimization in hIFN-γ production	Medium components (sorbitol, glycine, Na2HPO4, and MgSO4.7H2O)	Achieving maximum productivity in shake flasks level and bioreactor level	[101]
Pichia pastoris	MLP3 neural network was used to optimize the controller for a novel fed-batch production of A1AT	Controller gain and control poles	The cell growth and protein titer were improved significantly in comparison to traditional approaches	[102]
Rhodotorula glutinis	Establishing ANN and SVM models for microbial lipid production	Biomass, lipid yield, and COD removal rate	SVM performed better than ANN for small samples Fermentation parameters were optimized by integration of SVM and GA	[103]
Sludge	Developing a hybrid model for optimizing CBP production using LS-SVM and OED	Corn stalk weight, ultrasonic duration time, pretreatment time, and dual-frequency ultrasound	The model increases CBP 14.13% more than pure OED	[104]
Escherichia coli	Using random forest and ANN for biomass and recombinant protein modeling in a fed-batch process	Temperature, induction strength, growth rate, soluble/insoluble product formation	The accuracy reached about ±4% for dry cell mass and ±12% for protein concentration	[106]
saccharomyces cerevisiae	Employing GPR for prediction of *S. cerevisiae* biomass concentration	Molasses feed rate	Experimentally validation results showed high accuracy	[107]
Penicillium brevicompactum	Establishing a *k*-nearest-neighbor model to optimize different independent factors in MPA production	Ultrasound power, irradiation duration, treatment frequency and duty cycle	1.64-fold improvement observed in MPA production	[108]

Figure Legends:

Figure 1: Schematic of constraint-based modeling methods. FBA and MFA indicate metabolic flux distributions (fluxomics). A rational analysis of fluxomics causes the optimization of fermentation parameters, thus enhancing the product yield and reducing the production costs.

Figure 2: Machine learning workflow.

Figure 3: An overview of the structure of a dataset. This figure demonstrates a labeled dataset because one or more target values are reported. Datasets are usually prepared by data scientists. However, the more one knows about a dataset, the easier the machine learning process will be.

Figure 4: The role of cross-validation in machine learning. First, the generalization power of the ML model is evaluated through cross-validation. Then, hyperparameter tuning is performed before training the model to refine the model's parameters which are called hyperparameters. Grid search and Bayesian optimization algorithms are the most common search-based methods to tune hyperparameters. Finally, the curated model is used to evaluate the prediction capability of unseen test data.

Figure 5: Integrating constraint-based modeling and machine learning. CBM and ML can be integrated in different ways for analysis and optimization of fermentation parameters. **(a)** Predicting parameters by ML when fluxomics are used as inputs. **(b)**Predicting parameters by ML when the integration of fluxomics with multi-omics is used as input. **(c)** Predicting fluxomics and constraint-based models by ML when multi-omics are used as inputs.

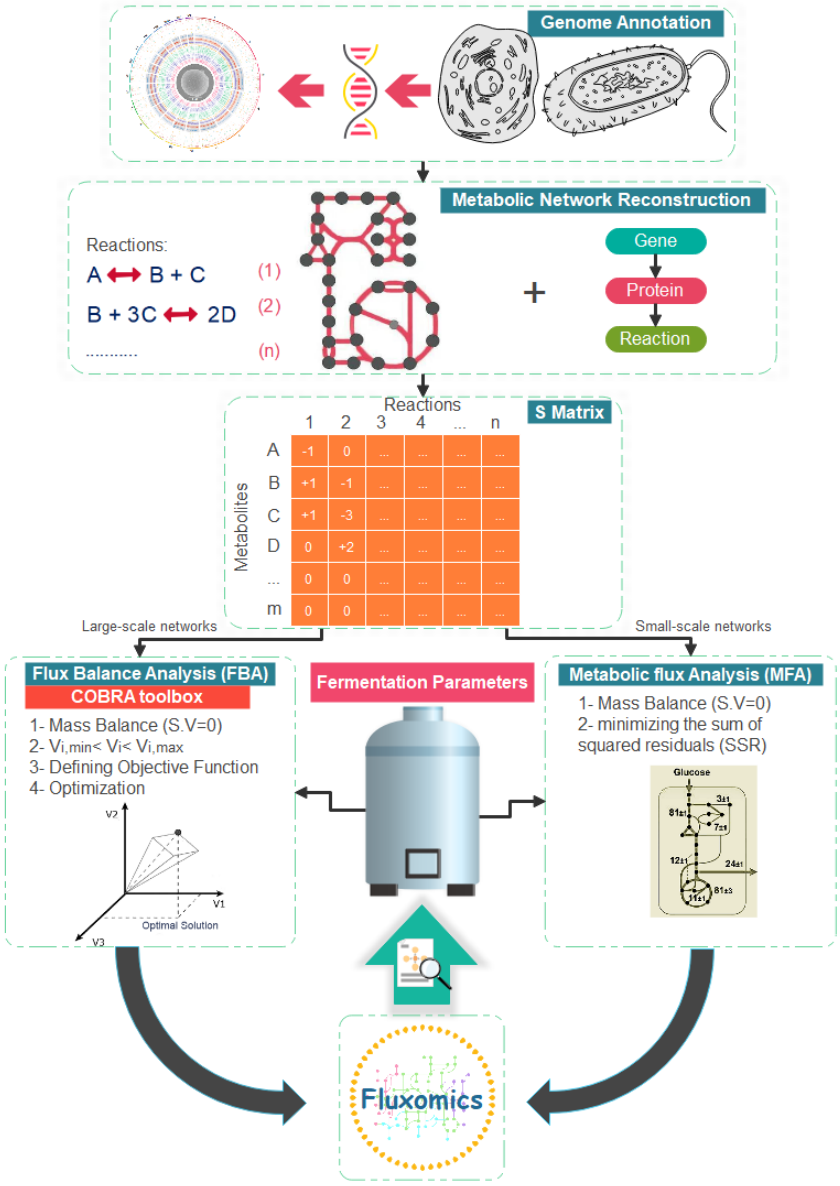

Figure 1

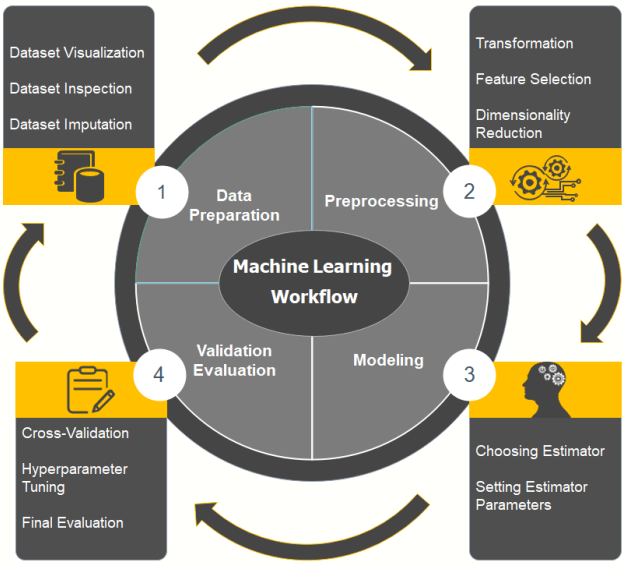

Figure 2

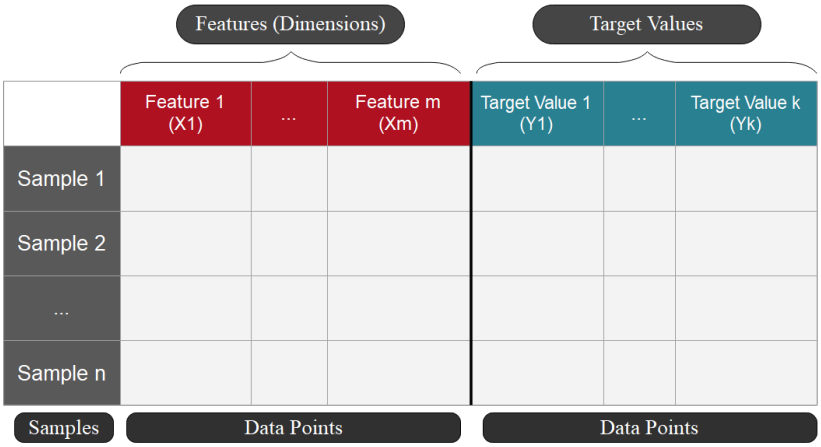

Figure 3

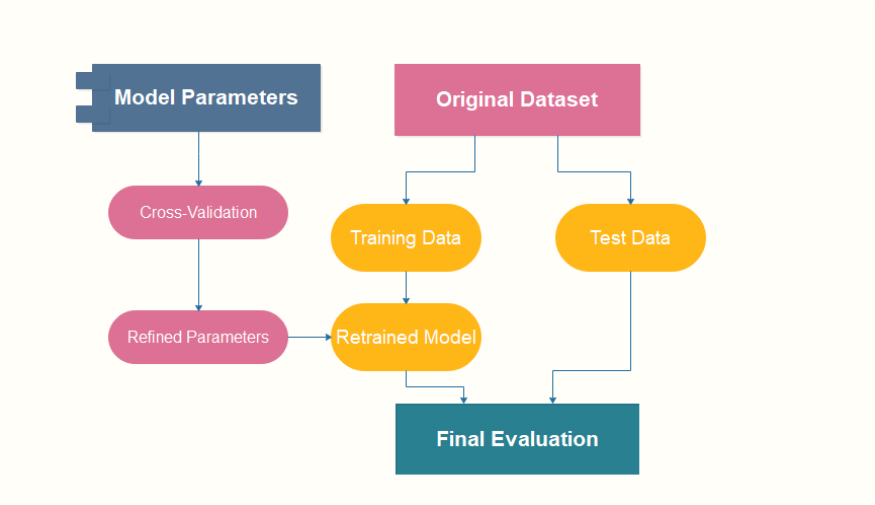

Figure 4

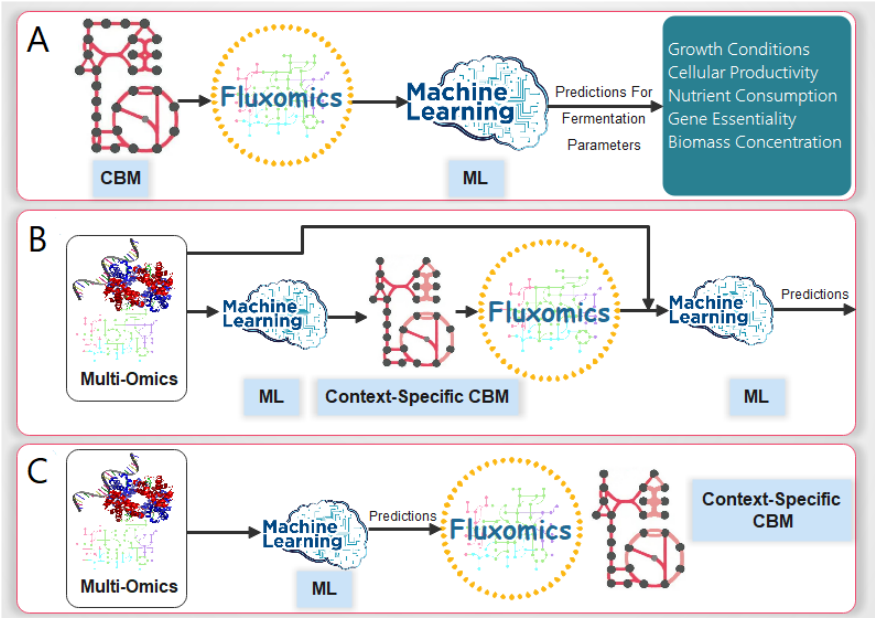

Figure 5